NPS-AM-08-041

EXCERPT FROM THE

PROCEEDINGS

OF THE

FIFTH ANNUAL ACQUISITION RESEARCH SYMPOSIUM

A NON-SIMULATION BASED METHOD FOR INDUCING PEARSON'S CORRELATION BETWEEN INPUT RANDOM VARIABLES

Published: 23 April 2008

by

Eric R. Druker, Richard L. Coleman and Peter J. Braxton

5th Annual Acquisition Research Symposium of the Naval Postgraduate School:

Acquisition Research: Creating Synergy for Informed Change

May 14-15, 2008

Prepared for: Naval Postgraduate School, Monterey, California 93943

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
23 APR 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Non-Simulation Based Method for Inducing Pearson's Correlation Between Input Random Variables	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Northrop Grumman IT (TASC),4584 Emerald View Ct,Eureka,MO,63025**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
5th Annual Acquisition Research Symposium: Creating Synergy for Informed Change, May 14-15, 2008 in Monterey, CA

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**45**	

The research presented at the symposium was supported by the Acquisition Chair of the Graduate School of Business & Public Policy at the Naval Postgraduate School.

To request Defense Acquisition Research or to become a research sponsor, please contact:

NPS Acquisition Research Program
Attn: James B. Greene, RADM, USN, (Ret)
Acquisition Chair
Graduate School of Business and Public Policy
Naval Postgraduate School
555 Dyer Road, Room 332
Monterey, CA 93943-5103
Tel: (831) 656-2092
Fax: (831) 656-2253
E-mail: jbgreene@nps.edu

Copies of the Acquisition Sponsored Research Reports may be printed from our website www.acquisitionresearch.org

Conference Website:
www.researchsymposium.org

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Proceedings of the Annual Acquisition Research Program

The following article is taken as an excerpt from the proceedings of the annual Acquisition Research Program. This annual event showcases the research projects funded through the Acquisition Research Program at the Graduate School of Business and Public Policy at the Naval Postgraduate School. Featuring keynote speakers, plenary panels, multiple panel sessions, a student research poster show and social events, the Annual Acquisition Research Symposium offers a candid environment where high-ranking Department of Defense (DoD) officials, industry officials, accomplished faculty and military students are encouraged to collaborate on finding applicable solutions to the challenges facing acquisition policies and processes within the DoD today. By jointly and publicly questioning the norms of industry and academia, the resulting research benefits from myriad perspectives and collaborations which can identify better solutions and practices in acquisition, contract, financial, logistics and program management.

For further information regarding the Acquisition Research Program, electronic copies of additional research, or to learn more about becoming a sponsor, please visit our program website at:

www.acquistionresearch.org

For further information on or to register for the next Acquisition Research Symposium during the third week of May, please visit our conference website at:

www.researchsymposium.org

THIS PAGE INTENTIONALLY LEFT BLANK

A Non-simulation Based Method for Inducing Pearson's Correlation between Input Random Variables

Presenter: Eric R. Druker graduated from the College of William and Mary with a BS in Applied Mathematics in 2005, concentrating in both Operations Research and Probability & Statistics with a minor in Economics. He is employed by Northrop Grumman as a Technical & Research lead. He performs cost and risk analysis on several programs within both the Intelligence and DoD communities. He was a recipient of the 2005 NGIT President's Award for his work on Independent Cost Evaluations, during which he helped develop the risk process currently used by NGIT's ICE teams. As a member of Northrop Grumman's ICE working group, he has helped shape the cost and risk practices used on independent cost estimates and evaluations across the corporation. In addition to SCEA conferences, Druker has also presented papers at the Naval Postgraduate School's Acquisition Research Symposium, DoDCAS and the NASA PM Challenge. He has also performed decision tree analysis for NG Corporate law and built models for Hurricane Katrina Impact Studies and Schedule/Cost Growth determination.

Eric R. Druker
Technical/Research Lead—Northrop Grumman IT (TASC)
4584 Emerald View Ct.
Eureka, MO 63025
Office: (636) 587-2624
Mobile: (703) 408-0589
Email: Eric.Druker@ngc.com

Author: Richard L. Coleman is a 1968 Naval Academy graduate. He received an MS with Distinction from the US Naval Postgraduate School and retired from active duty as a Captain, USN, in 1993. His service included tours as Commanding Officer of USS Dewey (DDG 45), and as Director, Naval Center for Cost Analysis. He has worked extensively in cost, CAIV, and risk for the Missile Defence Agency (MDA), Navy ARO, the intelligence community, NAVAIR, and the DD(X) Design Agent team. He has supported numerous ship programs including DD(X), the DDG 51 class, Deepwater, LHD 8 and LHA 6, the LPD 17 class, Virginia class submarines, CNN 77, and CVN 21. Coleman is the Director of the Cost and Price Analysis Center of Excellence and conducts Independent Cost Evaluations on Northrop Grumman programs. He has more than 65 professional papers to his credit, including five ISPA/SCEA and SCEA Best Paper Awards and two ADoDCAS Outstanding Contributed Papers. He was a senior reviewer for all the SCEA CostPROF modules and lead author of the Risk Module. He has served as Regional and National Vice President of SCEA and is currently a board member.

Richard L. Coleman
Sector Director—Northrop Grumman Cost/Price Analysis Center of Excellence
15036 Conference Center Dr.
Chantilly, VA 20151
Office: (703) 449-3627
Mobile: (703) 615-4482
Email: Richard.Coleman@ngc.com

Author: Peter J. Braxton holds an AB in Mathematics from Princeton University and an MS in Applied Science (Operations Research) from the College of William and Mary. He has worked to advance the state of knowledge of cost estimating, Cost As an Independent Variable (CAIV), Target Costing, and risk analysis on behalf of the Navy Acquisition Reform Office (ARO), the DD(X) development program, and other ship and intelligence community programs. He has co-authored several professional papers, including ISPA/SCEA International Conference award-winners in CAIV (1999) and Management (2005). Braxton served as managing editor for the original development of the acclaimed Cost Programmed Review of Fundamentals (CostPROF) body of knowledge and training course materials and is currently undertaking to lead a large team of cost professionals in a comprehensive update thereof. He serves as SCEA's Director of Training and was recently appointed a Northrop Grumman Technical Fellow.

Peter J. Braxton
Technical Fellow—Northrop Grumman IT (TASC)
15036 Conference Center Dr.
Chantilly, VA 20151
Office: (703) 961-3411
Mobile: (703) 944-3114
Email: Peter.Braxton@ngc.com

Abstract

Several previously published papers have cited the need to include correlation in risk-analysis models. In particular, a landmark paper published by Philip Lurie and Matthew Goldberg presented a methodology for inducing Pearson's correlation between input/independent random variables. The one subject, absent from the paper, was a methodology for finding the optimal applied correlation matrix given a desired outcome correlation. Since the publishing of the Lurie-Goldberg paper, there has been continuing discussion on its implementation; however, there has not been any presentation of an optimization algorithm that does not involve the use of computing-heavy simulations. This paper reviews the general methodology used by Lurie and Goldberg (along with its predecessor papers) and presents a non-simulation approach to finding the optimal input correlation matrix, given a set of marginal distributions and a desired correlation matrix.

Introduction

The Complete Correlation Algorithm (CCA) developed by Northrop Grumman and recently implemented in NG developed risk models is a product of more than two years of research and development. Several previously published papers have cited the need to include correlation in risk-analysis models; however, none present an optimization algorithm that does not involve the use of computing-heavy simulations. In particular, a landmark paper published by Philip Lurie and Matthew Goldberg (1998) presented a methodology for inducing Pearson's correlation between input random variables. This paper reviews the general methodology used by Lurie and Goldberg (along with its predecessor papers) and presents the Druker Algorithm: a non-simulation approach to finding the optimal input correlation matrix given a set of marginal distributions and a desired correlation matrix.

The *CCA* was deliberately created bearing in mind identified environmental factors that prevent easy implementation of commercially available models. No one on the team had any experience implementing correlation into Monte Carlo simulations beyond the use of COTS programs, such as @Risk™ and Crystal Ball™. To determine the best development method, the following factors were considered:

1. The Northrop Grumman risk models need to be of an easily transferable electronic size, as the models are often shared via email or network drives.
2. A diverse group of users must be able to run the software in a variety of work environments; Microsoft Office is the only platform that is transferable to all parties. Users include risk practitioners, program managers and members of pricing organizations; locations include unclassified and classified Northrop Grumman facilities, unclassified and classified customer facilities and home offices.

3. Custom implementations are frequent; much of NGIT-TASC risk work requires risk simulations to be built into pre-existing cost and price models. These models are generally limited to Microsoft Excel and Access; however, Web-based platforms are not unheard of.

The above concerns drove the decision to use Visual Basic source code to develop *the CCA.*

Initially, the development was focused on an algorithm that could induce Pearson's correlation between typical distributions in risk analysis: Bernoulli (discrete), Triangular, Normal and Log-Normal. By limiting the problem to the most-common applications, in theory, the solution should have been easier to find. While attempting to ascertain the maximum correlation between any two Bernoulli distributions, however, the general solution was uncovered. The resulting algorithm induces Pearson's correlation between any set of random variables (while still preserving the marginal distributions) using the Lurie-Goldberg Method and without the use of simulation to find the optimal applied correlation matrix.

The *CCA* is a compilation of multiple algorithms (each named for their main author(s)) from several sources: existing papers, public source code and internally-developed code. Most of the algorithms used were taken from a variety of existing papers. Although these papers all provided complete algorithms, they sometimes lacked details in how to accomplish key steps; in cases such as these, gaps were filled with open-source code solutions**.** The optimization of the applied correlation matrix, the last step in the correlation algorithm, was developed entirely by the Northrop Grumman Team.

Definitions and Assumptions

Matrix Definitions:

1. **Consistent Correlation Matrix—**Consistent Correlation matrices have diagonal entries equal to 1.0, all other entries between [-1, 1] are symmetric and positive definite. Consistency is necessary for a viable correlation matrix, but a Consistent Correlation Matrix may not necessarily be viable given the **Parent Distributions.**
2. **Input Correlation Matrix (I)—**The user-inputted correlation matrix. This matrix may or may not be a consistent correlation matrix.
3. **Adjusted Correlation Matrix (L)—**The **Input Correlation Matrix** adjusted to be a **Consistent Correlation Matrix.** This matrix will, by definition, be positive definite. Additionally, the adjusted matrix will be viable as correlations between various distributions of random variables will be achievable. When **(L)** is generated, the differences between **(I)** and **(L)** are minimized.
4. **Applied Correlation Matrix (A)—**The correlation matrix used by the *grand algorithm* to generate correlated random number draws. This matrix may be the same, or very different from, the **Adjusted Correlation Matrix;** the extent of the differences will depend on the random variables to be correlated.
5. **Optimal Applied Correlation Matrix (A')—**The Applied Correlation Matrix optimized using the Lurie-Goldberg Method.
6. **Outcome Correlation Matrix (O)—**The correlation matrix of the simulated variables following the simulation run. The goal of the *grant correlation algorithm* is for (O) to be identical to (L).

Other Definitions:

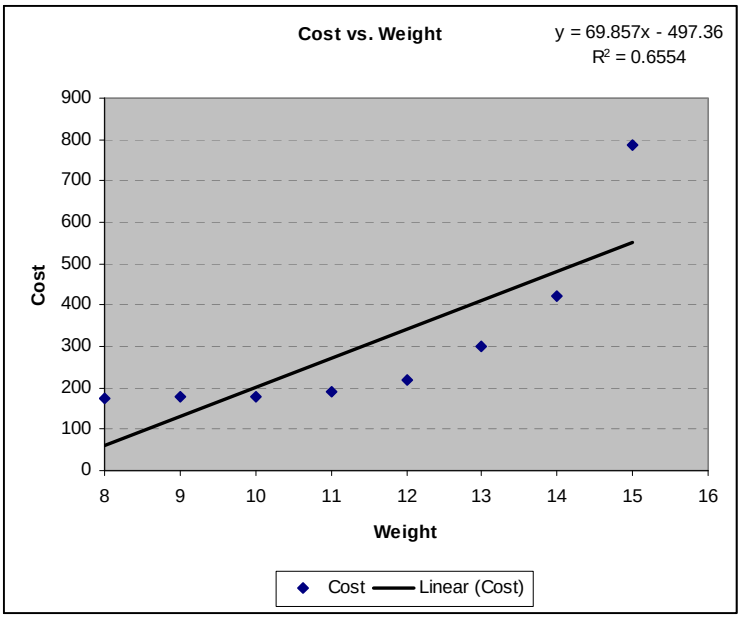
1. **Parent Distribution—**The distributions correlated for use in the simulation. The distributions are simulated using the Inverse CDF technique. The goal is to induce a desired correlation between these distributions.
2. **Pearson's Correlation—**A parametric statistic that measures the *strength and direction of a linear relationship between two random variables* ("Correlation," 2008).
3. **Spearman's Rank Correlation—**A non-parametric statistic that measures the monotonicity of a function without making any assumptions as to the distribution of the variables.
4. **Eigenvalues—**A scalar (L) associated with a matrix such that if (A) is a matrix and (X) is a vector, AX = LX. The vector (X) is known as the **Eigenvector** that corresponds to the **Eigenvalue** (L).

Assumptions:

1. **Normal Distributions—**Any reference to the normal distribution, whether in a univariate or bivariate case, is assumed to be the Standard Normal distribution (Mean of 0, Standard Deviation of 1).

Pearson's vs. Rank Correlation

Most COTS risk tools use Spearman's rank correlation as a substitute for Pearson's correlation between parent distributions. Spearman's rank correlation (a non-parametric statistic) differs from Pearson's correlation (a parametric statistic) in that it measures the monotony of a function, whereas Pearson's correlation measures the strength of the linear relationship between two functions (see Figure 1). Though studies have shown that, using the most common risk distributions, models using rank correlation yield similar results to those using Pearson's (Robinson & Salls, 2004), there is a distinct difference between the two. Although this paper will not detail all the differences between the two measures, a quick (and exaggerated) example is presented below. The *grand algorithm* supersedes the need to substitute for Pearson's correlation with Spearman's rank correlation.



Pearson's Rho	**0.81**
Spearman's Rho	**1.00**

Figure 1. Pearson's vs. Spearman's Rank Correlation

Algorithm Overview

There are three main steps behind the *grand algorithm.* An outline of these steps follows, and the upcoming sections of this paper will review each individual step in detail.

1. Correct the **User-Input Correlation Matrix (I)**
    a. Correct **I** so that it is consistent—both in terms of a general correlation matrix and the properties of the parent distributions being correlated.
    b. Through these corrections, the Adjusted Correlation Matrix **(L)** will be generated.
2. Optimize the **Applied Correlation Matrix**
    a. Find the **Optimal Applied Correlation Matrix (A')** such that when **A'** is run through the Lurie-Goldberg Method, the **Outcome Correlation Matrix (O)** is identical to **L**.
3. Correlate the Input Random Variables
    a. Using **A'**, apply the Lurie-Goldberg Method to correlate the parent distributions.

For purposes of presenting the methodology, it is necessary to show how the input random variables are to be correlated before discussing how to find **A'**.

Correcting the User-Input Correlation Matrix (Part I)

Giving users the ability to input their own correlation matrix allows for the possibility that the **User-Input Correlation Matrix (I)** may not be a viable correlation matrix. Correlation matrices, by definition, have diagonal entries of 1.0. All other entries between [-1, 1] are

symmetric and are positive definite. The first step in inducing correlation between input random variables is checking whether **I** is a consistent correlation matrix. If it is not, it must be corrected that it is such.

The *Iman-Davenport Algorithm*, which is based on a paper by Ronald Iman and James Davenport (1982) is used to correct **I** in order to make it a consistent correlation matrix. While numerous other papers have been published describing methods to correct **I** such that it is altered as little as possible (Higham, 2002), the *Iman-Davenport Algorithm* is the most computationally efficient method the authors uncovered. Given that additional adjustment may be required based on the parent distributions being correlated; the resulting matrix is close enough to **I** to satisfy this requirement.

The algorithm corrects **I** in three main phases. First, the algorithm checks whether **I** is symmetric with diagonal entries of 1.0 and off-diagonal entries between [-1, 1]. If it is not, the user is prompted to re-input the matrix, correcting for the discrepancies.

Second, once the above conditions are satisfied, the algorithm checks whether **I** is positive-definite. One way to test this is to find the eigenvalues for **I** (positive-definite matrices have all positive eigenvalues). The paper referenced did not describe an approach for finding the eigenvalues of the matrix. After further research, the Jacobi Eigenvalue Algorithm was determined to be a sufficiently efficient way to evaluate a matrix's eigenvalues. As a result, the eigenvalues are produced as the diagonals of an otherwise zero-matrix. The Jacobi Eigenvalue Algorithm is computationally inexpensive and pre-existing source code was used in its implementation.

If all eigenvalues for **I** are positive and the other conditions have been satisfied, then **I** is a consistent correlation matrix. Otherwise, in the third phase, negative eigenvalues are changed to small, positive values (e.g., .000001). The diagonal matrix of adjusted eigenvalues is then multiplied by the associated matrix of eigenvectors (also produced using the Jacobi Eigenvalue Algorithm). That product is, in turn, multiplied by the inverse of the matrix of eigenvectors to arrive at a new matrix that is similar, but not equal to, **I**. Lastly, the diagonals are reset to 1.0 as they may have changed during the transformation. This third section of the algorithm is repeated until all eigenvectors of the adjusted matrix are positive. At this point, the user input matrix has been adjusted such that it is a consistent correlation matrix.

Though the **User-Input Correlation Matrix** is now a consistent correlation matrix, the transformation of **I** is not complete and the **Adjusted Correlation Matrix (L)** has not been determined. As will be shown later, depending on the parent distributions being correlated, there may be a maximum achievable correlation between any two of the variables. Determination of **L** will be covered later in the section: *Correcting the User-Input Correlation Matrix (Part II)*.

Correlating Input Random Variables

In order to understand how the **Applied Correlation Matrix (A)** is to be optimized such that the **Output Correlation Matrix (O)** is identical to the **Adjusted Correlation Matrix (L)**, the method for correlating the parent distributions must first be discussed. It is a well-known fact that normal random variables can be correlated by multiplying a vector of uncorrelated normal

random draws by the Cholesky decomposition[1] of the desired correlation matrix. The Lurie-Goldberg Method takes this one step further using normal random variates to generate correlated uniform random variates. These uniform random variates are then transformed via the inverse-CDF technique to generate draws from the desired parent distributions. In this method, although the correlations between the normal random draws are known, as these draws are transformed into other distributions, the correlations change. Hence, the core problem emerges: how can the **Optimal Applied Correlation Matrix (A')** be uncovered such that **O** matches **L**? Answering this question is key to implementing the Lurie-Goldberg Method. The authors have developed an algorithm that addresses this very question, without necessitating any runs of the simulation. Additionally, they have begun the process of optimizing this algorithm, finding heuristics that allow it to run with a minimal number of calculations.

Implementation and Application of the *CCA*

The *CCA's* chief advantage is that it is non-recurring and its implementation requires no simulation. Furthermore, because the algorithm only requires looking at pairs of parent distributions, once the applied matrix has been found for a set of parent distributions, the algorithm must only be run when distributions are added or changed, and even then, only for the new/altered distributions. The algorithm also uses Pearson's correlation while COTS risk tools substitute Spearman's rank correlation.

The applications of the *CCA* reach beyond the Cost and Risk analysis community; this algorithm is useful anywhere there is a need to induce Pearson's correlation between input variables. For example, this algorithm can applied to auto correlating, stock market projections in the financial arena and to traditional modeling and simulation situations when correlation is needed. The algorithm was designed with a focus on portability. Because algorithm is coded with Visual Basic, it can be easily integrated in existing tools and models.

List of References

Correlation. (2008). *Wikipedia*. Retrieved April 1, 2008, from http://en.wikipedia.org/wiki/Correlation

Goldberg, M.S., & Lurie, P.M. (1998, February). An approximate method for sampling correlated random variables from partially-specified distributions. *Management Science*, *44*(2).

Higham, N. (2002, July). Computing the nearest correlation matrix—A problem from finance*. IMA Journal of Numerical Analysis*, *22*, 329-343.

Iman, R., & Davenport J. (1982, June). *An iterative algorithm to produce a positive definite matrix from an "approximated correlation matrix" (With a program user's guide).* Albuquerque, NM: Sandia National Laboratories.

Robinson, M., & Salls, W. (2004, June). *More on correlation accuracy in crystal ball simulations (or what we've now learned about Spearman's R in cost risk analyses)*. Remarks delivered at the 2004 SCEA Conference, Manhattan Beach, CA.

Open Source Code References:

The Foxes Team, Italy—http://digilander.libero.it/foxes

Axel Vogt, Germany—http://www.axelvogt.de/axalom/bivariateNormal_Series.zip

[1] The Cholesky Decomposition Matrix of any matrix M is L such that $M = LL^T$

THIS PAGE INTENTIONALLY LEFT BLANK

2003 - 2008 Sponsored Research Topics

Acquisition Management

- Software Requirements for OA
- Managing Services Supply Chain
- Acquiring Combat Capability via Public-Private Partnerships (PPPs)
- Knowledge Value Added (KVA) + Real Options (RO) Applied to Shipyard Planning Processes
- Portfolio Optimization via KVA + RO
- MOSA Contracting Implications
- Strategy for Defense Acquisition Research
- Spiral Development
- BCA: Contractor vs. Organic Growth

Contract Management

- USAF IT Commodity Council
- Contractors in 21st Century Combat Zone
- Joint Contingency Contracting
- Navy Contract Writing Guide
- Commodity Sourcing Strategies
- Past Performance in Source Selection
- USMC Contingency Contracting
- Transforming DoD Contract Closeout
- Model for Optimizing Contingency Contracting Planning and Execution

Financial Management

- PPPs and Government Financing
- Energy Saving Contracts/DoD Mobile Assets
- Capital Budgeting for DoD
- Financing DoD Budget via PPPs
- ROI of Information Warfare Systems
- Acquisitions via leasing: MPS case
- Special Termination Liability in MDAPs

Human Resources

- Learning Management Systems
- Tuition Assistance
- Retention
- Indefinite Reenlistment
- Individual Augmentation

Logistics Management

- R-TOC Aegis Microwave Power Tubes
- Privatization-NOSL/NAWCI
- Army LOG MOD
- PBL (4)
- Contractors Supporting Military Operations
- RFID (4)
- Strategic Sourcing
- ASDS Product Support Analysis
- Analysis of LAV Depot Maintenance
- Diffusion/Variability on Vendor Performance Evaluation
- Optimizing CIWS Lifecycle Support (LCS)

Program Management

- Building Collaborative Capacity
- Knowledge, Responsibilities and Decision Rights in MDAPs
- KVA Applied to Aegis and SSDS
- Business Process Reengineering (BPR) for LCS Mission Module Acquisition
- Terminating Your Own Program
- Collaborative IT Tools Leveraging Competence

A complete listing and electronic copies of published research are available on our website: www.acquisitionresearch.org

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CALIFORNIA 93943

www.acquisitionresearch.org

A Non-Simulation Based Method for Inducing Pearson's Correlation Between Input Random Variables

And its application on the CG(X) risk assessment

2008 Acquisition Research Symposium

15 May 2008

Eric Druker

Technical/Research Lead – Northrop Grumman IT

Acknowledgements

- Thanks to Dr. Steven Book of MCR for his help in obtaining copies of several papers on correlation modules, without which this paper would not have been complete
- Thanks to John Samberg of Tecolote for conversations that helped in the writing of this paper

Introduction

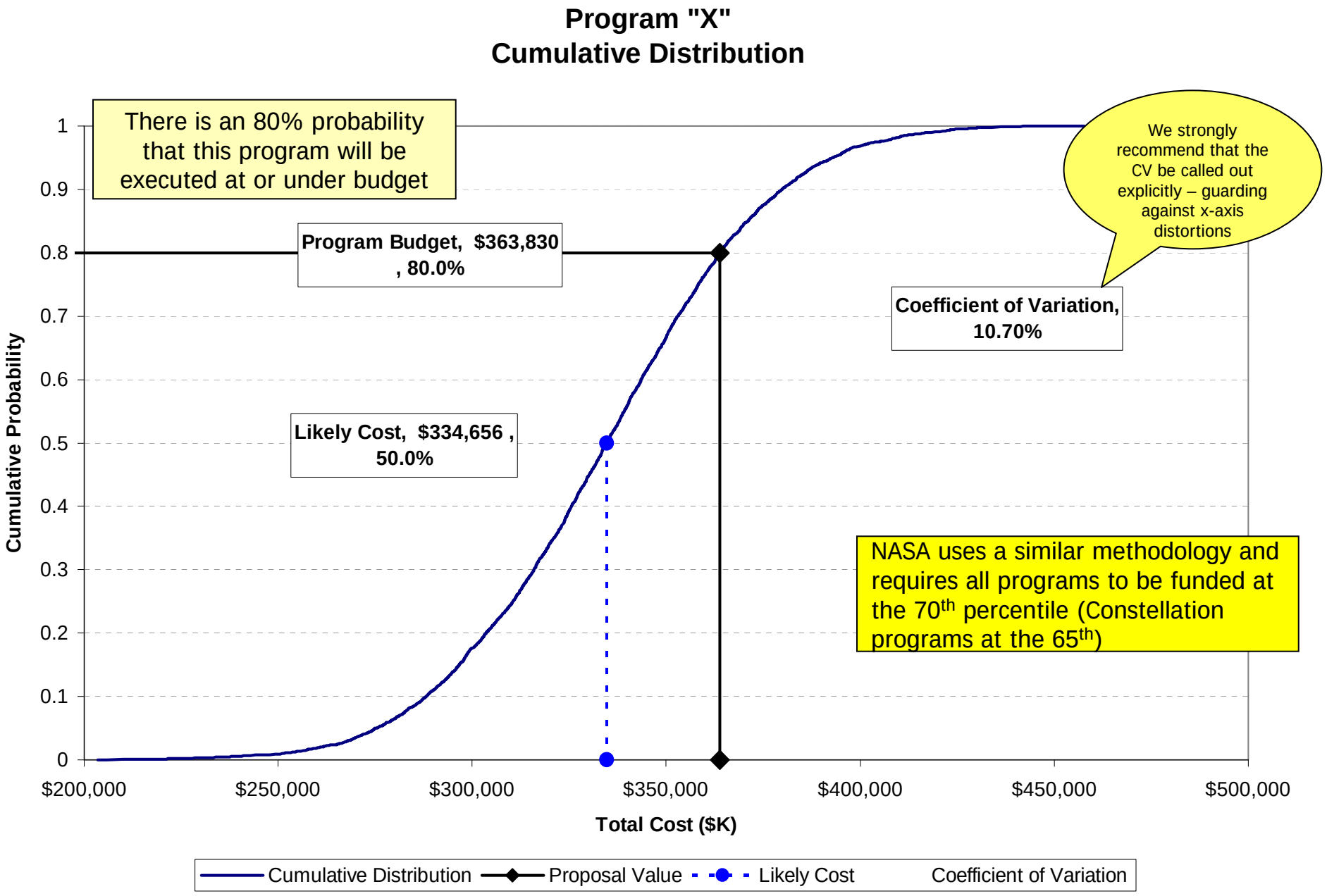
NORTHROP GRUMMAN

- **Before moving to the main topic of the paper it is important to quickly discuss the motivation behind its development**
- **Studies have shown[1, 2] that 75-85% of DoD programs experience cost overruns**
  - This suggests that as an industry, our estimates are not at the 50th percentile, but rather at about the 20th percentile
- **Recognizing this, agencies are taking the initiative to budget at higher percentiles of cost**
  - NASA requires all programs be funded at the 70th percentile
    - Constellation at the 65th
  - The Air Force (Dr. Sega) has released a memo advising that all space programs be funded at the 80th percentile
    - Rich Hartley (AFCAA) has advised against this, recommending programs be funded at the mean of the AFCAA ICE Estimate (generally between about the 55th and 60th percentiles)
- **In order to determine the appropriate funding level for programs anywhere but at the mean, it is thus imperative that the risk and uncertainty around estimates be assessed**
  - Thus S-Curves must be developed

1 Schaffer 2004 study, referenced from *Cost Estimating Requirements to Support New Congressional Reporting Requirements*. Coonce et. Al. NASA PM Challenge, February 2008

2 NAVAIR Cost Growth Study, R. L. Coleman, M.E. Dameron, C.L. Pullen, J.R. Summerville, D.M. Snead, 34th DoDCAS and ISPA/SCEA 2001

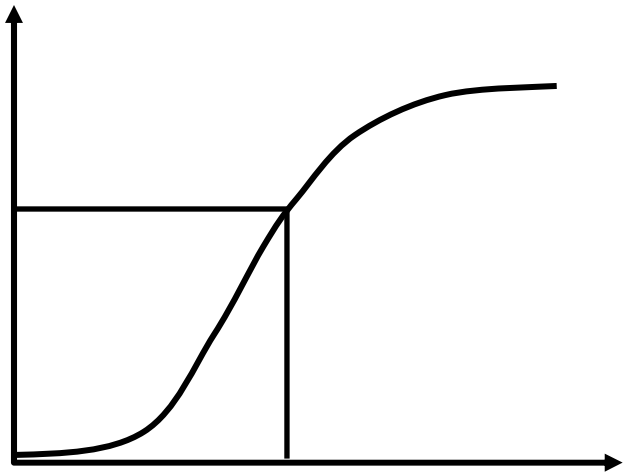
Sample Program S-Curve



S-Curves

NORTHROP GRUMMAN

- **S-Curves are the cumulative distribution function for the cost of a system**
  - Also known as probabilistic cost estimates
- **S-Curves are generally driven by two main factors**
  - Cost Estimating Variance
    - Labor estimates
      - Data Driven
      - SME Driven
    - Escalation/Inflation Rates
    - Material Costs
    - Productivity (e.g. hrs/SLOC, hrs/ft^2)
  - Schedule/Technical Risks and Opportunities
    - Discrete Events
    - Continuous Events



- **Two key measures are derived from these S-Curves**
  - Confidence level of the estimate
    - What is the probability that the program will finish at or under budget?
  - Uncertainty in the estimate
    - What is the range of possibilities for the final cost of this program?

Statement of Problem/Motivation

NORTHROP GRUMMAN

- Due to the increased focus on the reasonableness of cost estimates across the DoD community, a thorough risk assessment was conducted on the CG(X) program estimate
  - In particular, the Northrop Grumman team wanted to explore reasons that cost growth may be underestimated
  - It was determined that the treatment of correlation in risk adjusted cost estimates was one of the leading causes of this
  - Correlation directly effects the CV of the S-Curve
- In order to correctly capture program risk at a lower level, NGIT needed a way to include relational/injected correlation in our risk models
  - Without this ability the top level CV would be artificially shrunk due to the "square root of n problem"
- The following conditions lead the team away from traditional COTS models
  - The risk analysis module was to be incorporated into the CG(X) cost model
  - Both the cost and risk models were to be transitioned to a web-based platform
- In early 2006, work was begun on what would become the "Cost/Risk Correlation Module"
  - The module would have to exist entirely inside of Excel and VBA so it could be shared with any user with Office 2003 or later
  - The module would have to be *open* enough that it could be dropped quickly into most home-grown Monte Carlo models

Outline

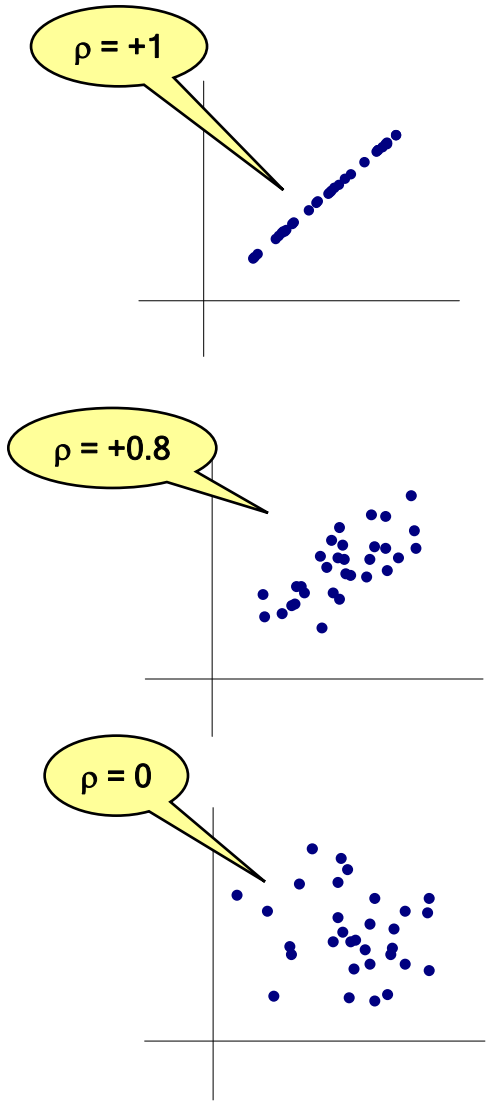
NORTHROP GRUMMAN

- Introduction to Correlation
  - Pearson's "Rho"
  - Pearson's vs. Rank Correlation
- The Problem
- Correlation Matrix Definitions
- Correlation in Risk Models
- Cost/Risk Correlation Algorithm
  - Correcting the user-input matrix
  - Correlating the random variables
  - Optimizing the Applied Matrix

Correlation (Pearson's)

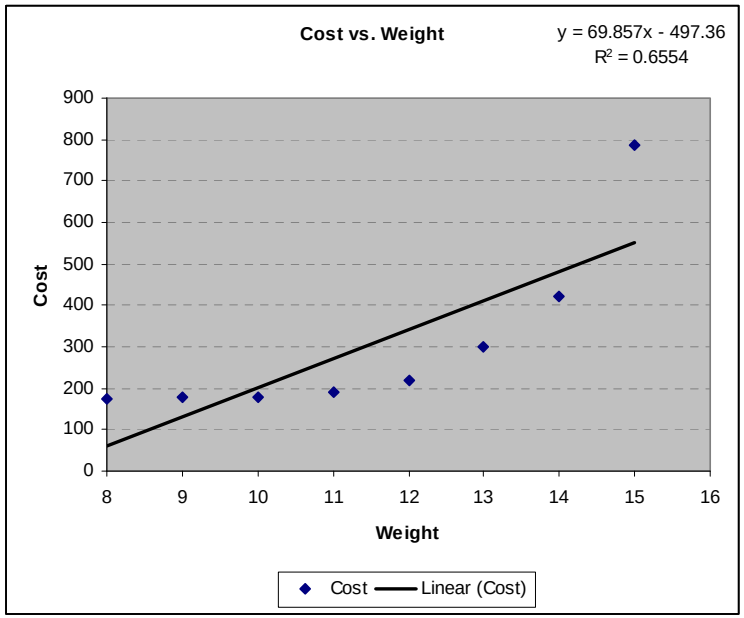
NORTHROP GRUMMAN

- Although this paper is not about correlation itself, it's important to briefly review the two most common measures
  - Pearson's Product-Moment Correlation
  - Spearman's Rank Correlation
- When correlation is discussed in terms of cost estimating, Pearson's correlation is generally described
- Pearson's Correlation is a measure of the linear relationship between two or more variables
  - This is as opposed to Rank Correlation, which will be discussed on the next slide
- It is identified using the Greek symbol ρ and is always between [-1,1]
- The correlation of a linear regression is the square root of r^2
- The examples on the right show representative data sets for three values of ρ



Pearson's Correlation vs. Rank Correlation

- Most commercial risk programs (e.g. Crystal Ball & @Risk) use Spearman's rank correlation rather than Pearson's correlation because it is easier to simulate
- Spearman's rank correlation is used to detect correlation between two variables, without assuming a linear relationship
  - It is concerned with whether or not the function is monotonic
- Some other differences include
  - Pearson's is parametric, Spearman's is not
  - Spearman's is not to be used for predictive purposes
- In the example to the right, rank correlation and Pearson's correlation yield very different answers
- Although it is important to distinguish between these two types of correlation, past research has shown that in cost risk simulations using the most common families of distributions, the two yield fairly similar results[1]
  - The aim of the authors is to "commit no avoidable errors"



Pearson's Rho	**0.81**
Spearman's Rho	**1.00**

[1] Robinson, M and Salls, W. *More on Correlation Accuracy in Crystal Ball Simulations (or What We've Now Learned about Spearman's R in Cost Risk Analyses)*. Presented at the 2004 SCEA Conference, Manhattan Beach, CA, June 2004

Correlation in Risk Models

NORTHROP GRUMMAN

- In risk analysis, correlations are critical to successful simulations used to find distributions of cost
  - Correlations are thought to be widely present among elements of cost, but little data exists to determine them, principally because to determine correlations among any set of variables, data points must contain those variables in common, and this is rarely the case
  - Without accounting for correlation, summing multiple independent risk distributions will lead to an artificial degradation in the CV
    - This is known as the “Square Root of N” problem
- Lacking discernable correlations, risk analysts are forced to rely on Subject Matter Experts to estimate correlations
  - These correlations are subtle and difficult to estimate
  - Estimated correlations, to be usable, must be “coherent”, as discussed later
- Once the desired correlation between all cost elements is determined, the next problem is to build these correlations into the risk model
- The following slides will lay out the algorithms used in the correlation module and demonstrate how they were applied to the CG(X) program

Definitions: Matrices

NORTHROP GRUMMAN

- Before proceeding, it is important to define several matrices that will be used in the algorithm
- Input Correlation Matrix:
  - The correlation matrix inputted by the user, may or may not be a consistent correlation matrix
- Adjusted Correlation Matrix:
  - The consistent correlation matrix found by the model that is as close as possible to the Input Correlation Matrix
    - This matrix is positive semidefinite
    - It is also coherent given the distributions being correlated
- Applied Correlation Matrix:
  - The correlation matrix utilized by the algorithm to generate correlated random number draws
- Outcome Correlation Matrix
  - The correlation matrix of the simulation variables after the simulation is run
  - Ideally it is identical to the Adjusted Correlation Matrix

User-Input Matrix		
1.0000	0.8000	0.1000
0.8000	1.0000	0.8000
0.1000	0.8000	1.0000

Adjusted Matrix		
1.0000	0.7522	0.1322
0.7522	1.0000	0.7522
0.1322	0.7522	1.0000

Applied Matrix		
1.0000	0.7915	0.2263
0.7915	1.0000	0.7744
0.2263	0.7744	1.0000

Outcome		
1.0000	0.7522	0.1316
0.7522	1.0000	0.7521
0.1316	0.7521	1.0000

Definitions: Eigenvalues/Eigenvectors

NORTHROP GRUMMAN

- An eigenvector is a vector v such that for a square matrix A and a scalar λ, $Av = \lambda v$
- It follows that if Q is an indexed set of linearly independent eigenvectors for matrix A and Λ is the diagonal matrix containing the corresponding eigenvalues of A as its diagonal entries then:

  $A = Q\Lambda Q^{-1}$

- By altering Λ, the diagonal matrix consisting of A's eigenvalues, we eventually arrive at a positive definite correlation matrix that is close to the user input matrix
- The Jacobi Eigenvalue algorithm is used to find both the eigenvalues and eigenvectors of the user input correlation matrix

NORTHROP GRUMMAN

The Cost Risk Correlation Algorithm

Correcting the User Input Matrix

Correlating the Uniform Random Number Draws

Optimizing the Applied Matrix

Correcting the User Input Matrix

NORTHROP GRUMMAN

- As a rule, correlation matrices must be positive semidefinite
  - Positive semidefinite matrices have all non-negative eigenvalues
- When using data to generate correlation matrices, they will necessarily be positive definite
- Unfortunately, when generating matrices based on SME judgment, this condition may not be met
- To correct these matrices, an algorithm developed by Iman and Davenport[1] was used
  - The criteria for "closest matrix" that comes out of this algorithm is unknown to the authors but it is computationally efficient and relatively simple to implement
  - Because the generation of the "closest viable correlation matrix" is so critical in finance, there are several more robust algorithms available[2]
- The following slide will outline the algorithm used in the Cost-Risk Correlation Module

[1] Iman, R and Davenport J. *An Intterative Algorithm to Produce a Positive Definite Matrix from an "Approximated Correlation Matrix" (With a Program User's Guide)* Sandia National Laboratories for the US DoE, June 1982

[2] Higham, N. *Computing the Nearest Correlation Matrix – A Problem from Finance.* IMA Journal of Numerical Analysis. 2002

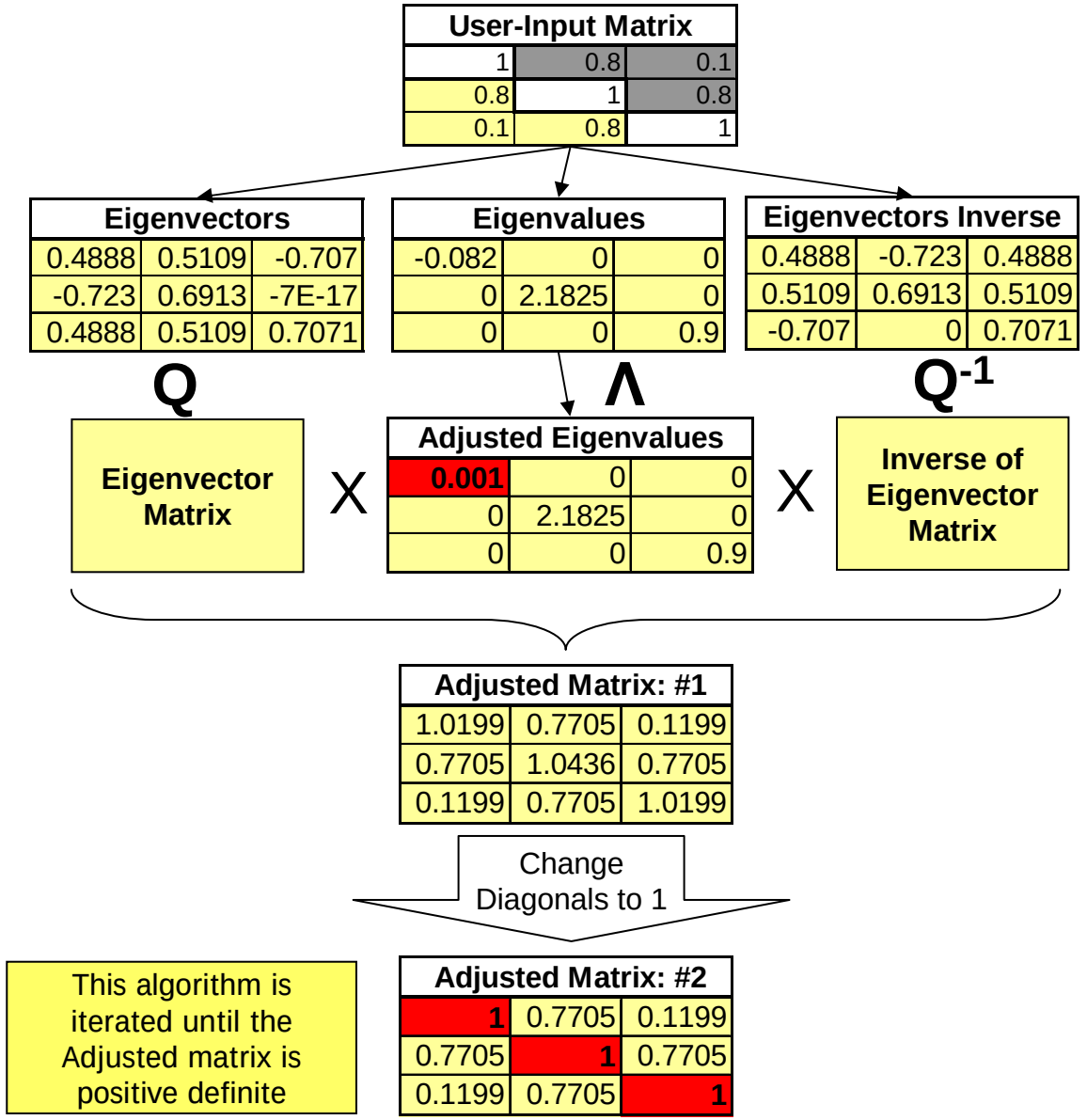
Correcting the User Input Matrix - Hurdles

NORTHROP GRUMMAN

- Two hurdles existed in implementing the algorithm
  - Excel doesn't have a function that finds Eigenvalues and Eigenvectors for the correlation matrices
  - Excel doesn't have a function to compute the Cholesky Decomposition matrix
- Research was conducted and algorithms (and the associated VBA source code) that conquered both hurdles were found
  - Both were part of the MATRIX and LINEAR ALGEBRA Package For EXCEL developed by The Foxes team in Italy
  - The Cholesky Decomposition, Eigenvalues and Eigenvectors functions were taken from this package and added into the tool

Correcting the User Input Matrix - Algorithm

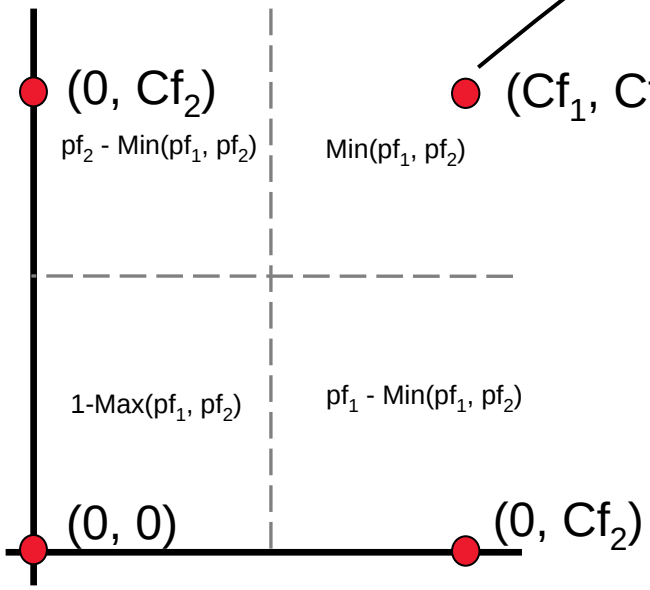
- The algorithm iteratively adjusts the eigenvalues of user-inputted correlation matrices until the resulting matrix has all non-negative
- During each iteration of the algorithm, there are two adjustments
  1. Adjustment of the negative eigenvalues to small, positive values
  2. Adjustment of the first adjusted matrix's diagonal entities to values of 1
- Once the adjusted matrix (#2) is found to have all non-negative Eigenvalues, the algorithm has found its solution



Correcting the User Input Matrix – Other Complications

- Although the matrix produced using the algorithm on the preceding slides is a consistent correlation matrix, depending on the random variables being correlated it may or may not be feasible
  - At least if the marginal distributions are to be preserved
- The best way to illustrate this is to examine the maximum possible correlation between two Bernoulli risks
  - As shown below, unless the probabilities of the two risks are equal, there is a maximum possible correlation between them
- The final step to correcting the User Input Matrix is to adjust the matrix so that all correlations are feasible based on the distributions being correlated

Although this example seems odd, this is an efficient way of inducing conditional probabilities between Bernoulli random variables



The only case in which XY ≠ 0 is when both risks occur, it follows that E(XY) simplifies down to $Cf_1 x Cf_2$ times the probability that both risks occur. The highest this probability can possibly be is the minimum of the two probabilities

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

$$\rho(\max)_{X,Y} = \frac{Min(Pf_1, Pf_2) \times Cf_1 \times Cf_2 - (Pf_1 \times Cf_1) \times (Pf_2 \times Cf_2)}{\sigma_X \times \sigma_Y}$$

$$\rho(\max)_{X,Y} = \frac{Min(Pf_1, Pf_2) \times Cf_1 \times Cf_2 - (Pf_1 \times Cf_1) \times (Pf_2 \times Cf_2)}{Cf_1\sqrt{Pf_1 \times Qf_1} \times Cf_2\sqrt{Pf_2 \times Qf_2}}$$

$$\rho(\max)_{X,Y} = \frac{Min(Pf_1, Pf_2) - Pf_1 \times Pf_2}{\sqrt{Pf_1 \times Qf_1}\sqrt{Pf_2 \times Qf_2}}$$

Correlating Random Variables: An Introduction to the Lurie-Goldberg Method[1]

NORTHROP GRUMMAN

- The only method the authors were aware of for inducing Pearson's correlation between input random variables is the Lurie-Goldberg Algorithm
  - The Lurie-Goldberg Algorithm aims to find an applied correlation matrix such that the input correlation and output correlation are as close as possible

- Find matrix **L** such that series of transformations

$$\underset{\text{indep. normal}}{X} \xrightarrow{\mathbf{L}} \underset{\text{mult. normal}}{Y} \xrightarrow{\Phi} \underset{\text{uniform}}{U} \xrightarrow{F^{-1}} \underset{\text{desired}}{V}$$

lead to random variables with desired correlations and marginal distributions

- **L**: Cholesky factor transforms independent normals to correlated normals
- Φ: normal c.d.f. transforms correlated normals to correlated uniforms
- F^{-1}: transforms correlated uniforms to correlated random variables with desired marginal distributions F

  - Unfortunately, the authors could not find a method for finding this optimal matrix (L... referenced as A' in this paper)
  - One obvious solution is to optimize the matrix by examining the post-simulation correlations
    - Given the computing power needed to complete each simulation, this could be a time consuming endeavor

[1]Goldberg, Matthew S, Lurie, Phillip M. *Correlating Random Variables*, 32nd DoDCAS, Williamsburg, VA. February 1999

Correlating the Uniform Draws: The Lurie-Goldberg Method

NORTHROP GRUMMAN

- Once a viable correlation matrix exists Uniform (0,1) correlated random numbers must be generated which in turn are used to generate the desired random variables
- To accomplish this, the Cholesky Decomposition Matrix of the adjusted matrix is found
  - L is the Cholesky Decomposition of A iff L is a lower triangular matrix such that:

$$A = LL^T$$

- After the Cholesky Decomposition Matrix is found, the algorithm at right is run to produce correlated Uniform (0,1) random numbers
- These random numbers, vice the originals, are used in the risk model to generate points off of distributions

Adjusted Matrix		
1.0000	0.7522	0.1322
0.7522	1.0000	0.7522
0.1322	0.7522	1.0000

Cholesky Decomposition

Cholesky Decomposition		
1.0000	0.0000	0.0000
0.7522	0.6589	0.0000
0.1322	0.9907	0.0321

X

U(0,1) Random Draws
0.26271853333989800
0.79616660202169400
0.15362541632109700

Inverse CDF Technique

Random N(0,1)
(0.63498673467686800)
0.82800654029771300
(1.02100761130346000)

Multiply N(0,1) by Cholesky

Correlated Random N(0,1)
(0.63498673467686800)
0.06794090908429620
0.70360627328862900

Multiply N(0,1) by Cholesky

Correlated Random U(0,1)
0.26271853333989800
0.52708366338494000
0.75916099823068700

Note: The resulting correlation between the Correlated Random U(0,1) random numbers will not be exactly the same as the adjusted correlation matrix… **more on this soon**

Outcome Correlation		
1.0000	0.7386	0.1367
0.7386	1.0000	0.7395
0.1367	0.7395	1.0000

Resulting Correlation

Optimizing the Applied Correlation Matrix

NORTHROP GRUMMAN

- Non-linear transformations are used to correlate random variables in the model
  - Because of this, the outcome correlation may be different from the intended correlation
- The biggest hurdle this module faced was in the correction of this discrepancy
- Northrop Grumman has developed a method that can find the outcome correlation matrix for any applied correlation matrix prior to the simulation being run
  - In other words, the algorithm can determine ρ_{Output} given $\rho_{Applied}$
  - The applied correlation matrix can then be optimized so that the outcome correlation matrix is equal to the adjusted correlation matrix
- Additionally, it follows from mathematical proofs that the optimal applied correlation matrix will induce the desired correlation
  - This infers that any variation in ρ in the simulation runs is due solely to Monte Carlo sampling error

Find:

Applied Correlation Matrix		
1.0000	0.7915	0.2263
0.7915	1.0000	0.7744
0.2263	0.7744	1.0000

Such that after the Lurie-Goldberg method takes place:

Outcome Correlation Matrix		
1.0000	0.7522	0.1322
0.7522	1.0000	0.7522
0.1322	0.7522	1.0000

=

Adjusted Correlation Matrix		
1.0000	0.7522	0.1322
0.7522	1.0000	0.7522
0.1322	0.7522	1.0000

Optimizing the Applied Correlation Matrix

NORTHROP GRUMMAN

- The algorithm developed by Northrop Grumman finds the optimal applied correlation matrix given:
  1. The parent distributions being correlated
  2. The adjusted correlation matrix
- The algorithm runs prior to the simulation being executed and once performed, only needs to be re-ran as variables are added or changed
  - And in those cases, only for the new/modified distributions
- Although the algorithm was originally developed for cost risk analysis, it has applications wherever a user needs to account for correlation between independent random variables
  - For example: the modeling of mutual fund performance given it is made up of a group of correlated stocks and bonds
- In fact, the algorithm's first use is in the modeling of conditional probabilities between Bernoulli independent random variables
  - The customer needed an efficient way to model the conditional probabilities they found between parameters in their data while preserving the marginal probabilities
  - It can be shown using the same general methodology on slide 14 that Pearson's correlation between two Bernoulli random variables equates to a conditional probability between them

Application to the CG(X) Program Risk Assessment

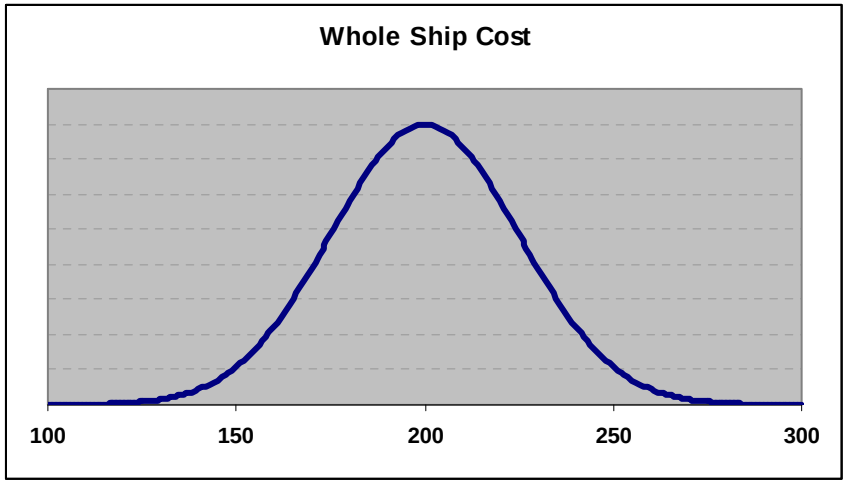
Correlation Data

NORTHROP GRUMMAN

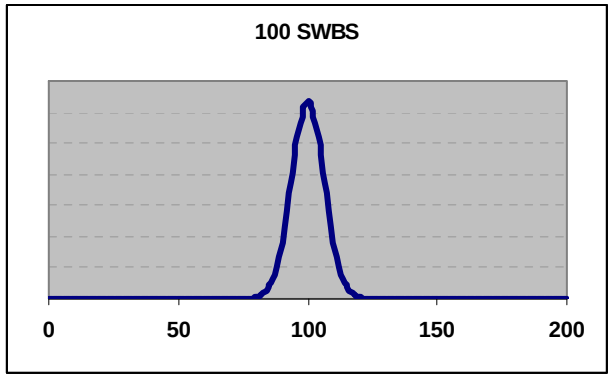
- One of the most difficult steps in the risk assessment process is in determining the correlation between the elements
- In this assessment, correlation is currently being measured using the relationship between the SWBS hours for three classes of surface combatants
- Just recently, data was obtained showing estimates vs. actuals, by SWBS, for various ships
  - The plan is to switch to correlations using this data once the analysis is complete
- Once uncertainty was evaluated for each lower level SWBS, correlation was applied between them to produce the top level risk adjusted estimate

Estimating Variance (diagram)

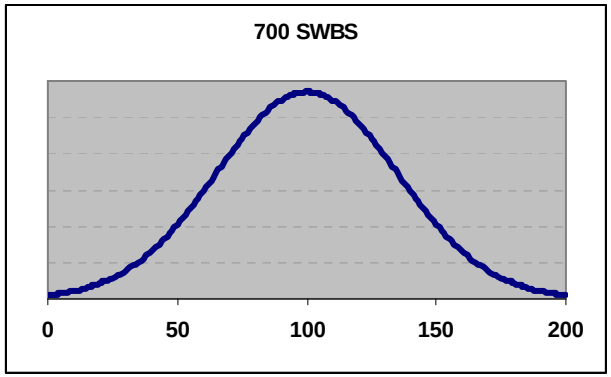
- This simplified example shows only the 100 and 700 SWBS
  - 100 may have a lower level of uncertainty around its estimate than 700
- Using the correlation algorithm, accurate distributions can be generated for the lower level SWBSs that, when added together, still produce the known historical distribution
  - This allows decision makers to see what areas of the ship contain the greatest variance
  - It also allows risk to be applied at the 1-digit-level (see next slides)



The bigger ratio of new to repeat work in the 700 SWBS is reflected in its larger CV (wider curve)



+



$\rho_{100,700}$

Schedule/Technical Risks & Opportunities

NORTHROP GRUMMAN

- The next step in the risk assessment was adding in schedule and technical risks
  - Opportunities are just risks with a negative cost impact (cost is decreased)
  - From this point on, risks &opportunities will be referred to simply as risks
- Technical experts involved in CG(X) from across the corporation were interviewed to produce the schedule/technical risks associated with their area of the ship
- The following information was collected:
  - Description of the risk
  - Probability of occurrence
  - Description of the impact
    - This is the consequence of the risk occurring
  - Mitigation plans
    - Description of the mitigation plan
    - Cost of the mitigation plan
    - Probability and impact if the risk is mitigated
    - Whether or not the mitigation plan is included in the cost baseline
  - Other areas of the ship affected if the risk were to occur
    - If a schedule/technical risk increased the probability of occurrence for another risk, this was captured using the previously described correlation algorithm

Schedule/Technical Risk Template

NORTHROP GRUMMAN

Risk ID:	An ID used to identify the risk. Label Sequentially
Risk Description:	The risk description is a basic description of what the risk is. In particular, what could go wrong.
Probability of Occurrence:	The probability that the risk will occur.
Impact Description:	The impact description is all the information that would be needed from the SME in order to estimate the cost impact of the risk independently. Wherever, possible, please include schedule impacts as well
Mitigation Plans(s):	The mitigation plan(s) are all activities that would lower the expected value of the risk. These activities do not have to completely eliminate the risks, they could just lower either the probability of occurrence or cost impact. Information to be included: 1. Cost of Mitigation Plan (both schedule and $) 2. Affect mitigation plan has on the risk (what is the decrease in probability or cost/schedule impact)
Other Areas Affected:	Are there any other areas of the ship that could be impacted if this risk were to occur (or if the mitigation plans are put into motion)? If so, describe the impact and the area it would affect. Then, interview the owner of that area to determine if there are anymore residual impacts not forseen originally.

Schedule/Technical Risk Modeling

NORTHROP GRUMMAN

- Once the risks are collected, they were input into the model
- For risks with mitigation strategies, whether or not the mitigation strategy is implemented was selected using a drop-down menu
  - Mitigated risks (whose cost of mitigation is not included in the cost baseline) will add cost to the baseline cost
  - Mitigated risks will use mitigated probabilities and consequences
- Each risk is assigned to a 1-digit-level SWBS
  - This, along with the fact that cost estimating variability is also assessed at the 1-digit-level, allows cost distributions to be produced accurately at the 1-digit-level
- Risks can also be inputted as continuous risks (as appropriate):
  - Triangular Distributions
  - Normal Distributions
  - Log-Normal Distributions
  - All of these distributions can have probabilities assigned to them as well

Schedule/Technical Risks

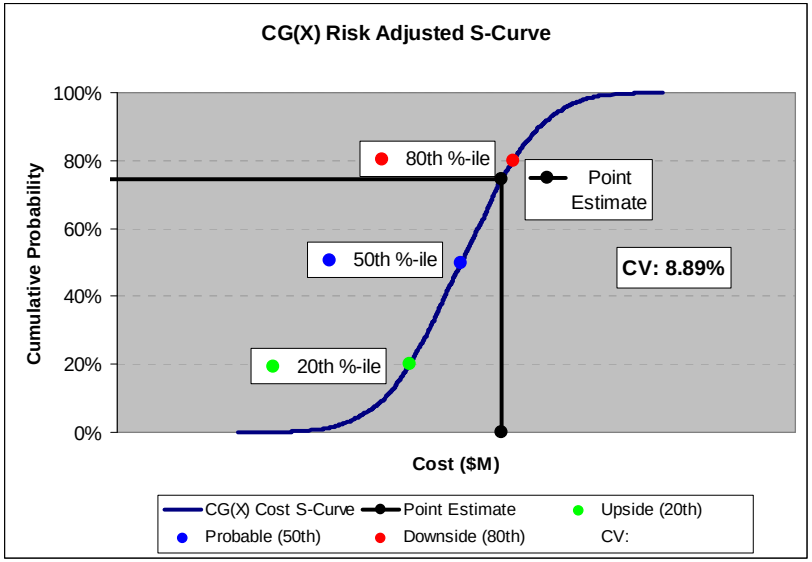
NORTHROP GRUMMAN

Risk ID	SWBS	Description	Probability of Occurrence	Bernoulli Cost Impact	Mitigation Plan Description	Mitigated Probability	Mitigated Cost Impact	Cost of Mitigation	Mitigation Plan Implemented ?
1	000	Sample Risk 1	90%	$ 25,000,000	Mitigation Plan 1	30%	$ 10,000,000	$ 7,500,000	Yes
2	200	Sample Risk 2	52%						No
3	300	Sample Risk 3	75%						No
4	400	Sample Risk 4	100%						No
5	500	Sample Risk 5	10%	$ 100,000,000	Mitigation Plan 5	1%	$ 50,000,000	$ 10,000,000	Yes
6	600	Sample Risk 6	25%	$ 13,000,000					No
7	700	Sample Risk 7	90%	$ 9,000,000					No
8	800	Sample Risk 8	100%						No
9	900	Sample Risk 9	100%						No

Model Also Accepts Triangular, Normal and Lognormal Risk Distributions

Results

- Several sets of results are produced automatically by the simulation when the “Run Simulation” button is hit
- CG(X) Risk Adjusted S-Curve
  - Shows the whole-ship cost distribution with the point estimate and its confidence on the graph
- CG(X) Risk Adjusted Estimate by 1-digit-level SWBS
  - Shows upside (20th Percentile), Probable (50th Percentile) and Downside (80th Percentile) by 1-digit-level SWBS
- CG(X) Risk by SWBS
  - Shows upside, probable and downside risk $’s by SWBS
  - These are the $’s due entirely to the risks, not estimating variation



CG(X) Risk Adjusted Estimate

SWBS	Description	Upside	Probable	Downside
000	Administration			
100	Hull			
200	Propulsion			
300	Electric Plant			
400	Electonics Systems			
500	Auxillary Systems			
600	Outfit & Furnishings			
700	Weapons			
800	Integration & Engineering			
900	Ship Assembly & Support			
	Total			

CG(X) Risk by SWBS

SWBS	Description	Upside	Probable	Downside
000	Administration			
100	Hull			
200	Propulsion			
300	Electric Plant			
400	Electonics Systems			
500	Auxillary Systems			
600	Outfit & Furnishings			
700	Weapons			
800	Integration & Engineering			
900	Ship Assembly & Support			
	Total			

Conclusion

NORTHROP GRUMMAN

- The previously discussed method is an attempt at producing a risk adjusted estimate for the CG(X) program that is also accurate at the SWBS level
- This analysis would not have been possible were it not for the creation of the cost/risk correlation module